0.7. Sums of Random Variables, Covariance, Correlation

Note. In this section, we consider a sum and linear combination of two random variables. For sums, we describe the joint probability function (in the discrete case) and the joint distribution function (in the continuous case). We find the variance of a linear combination of random variables and define covariance and correlation.

Note. Let (X, Y) be a discrete two-dimensional random vector. To find the probability function of the sum $X + Y$, we need the probabilities of the events $\{\omega \mid X(\omega) + Y(\omega) = x\}$ for all $z \in \mathbb{R}$. Now $X(\omega) + Y(\omega) = z$ precisely when $X(\omega) = x$ and $Y(\omega) = y$ where $x + y = z$, so

$$p_{X+Y}(z) = \sum\sum\nolimits_{\{(x,y)\mid x+y=z\}} p_{X,Y}(x,y) = \sum_x p_{X,Y}(x, z-x) \text{ for } z \in \mathbb{R}.$$

If, in addition, random variables X and Y are independent, then (by the definition from Section 0.6) we have $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ and so

$$p_{X+Y}(z) = \sum_x p_{X,Y}(x, z-y) = \sum_x p_X(x)p_Y(z-x) \text{ for } x \in \mathbb{R}.$$

But observes that in this "we recognize as the *convolution formula*" (page 9).

Note. In Mathematical Statistics 1, the convolution formula is addressed in the continuous setting, but only in an exercise (Exercise 2.2.5 in the text book) of Section 2.2. Transformations: Bivariate Random Variables. The exercise gives for the continuous case that the joint distribution function satisfies

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x)\,dx \text{ for } z \in \mathbb{R}.$$

If, in addition, X and Y are independent then

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,dx \text{ for } z \in \mathbb{R}.$$

Note. Expectation is linear: $E(aX + bY) = aE\ X + bE\ Y$ (or, more clearly, $E(aX+bY) = aE(X)+bE(Y)$). "It is easy to check" (Gut, page 10) and is covered in Mathematical Statistics 1 (STAT 4057/5057) in Section 1.8. Expectation of a Random Variable; notice Theorem 1.8.2. This lets us compute the variation of $aX + bY$:

$$\mathrm{Var}(aX + bY) = E[(aX + bY) - E(aX + vY)^2]$$

$$\begin{aligned}
&= E[(aX+bY)^2 - 2(aX+bY)E(aX+bY) + (E(aX+bY))^2] \\
&= E[a^2X^2 + 2abXY + b^2Y^2 - 2(aX+bY)(aE(x)+bE(Y)) + (aE(X)+bE(Y))^2] \\
&= E[a^2X^2 + b^2Y^2 + 2abXY - 2a^2XE(X) - 2abXE(Y) - 2abYE(X) - 2b^2YE(Y) \\
&\quad +a^2(E(X))^2 + 2abE(X)E(Y) + b^2(E(Y))^2] \\
&= E[a^2(X^2 - 2XE(X) + (E(X))^2) + b^2(Y^2 - 2YE(Y) + (E(Y))^2) \\
&\quad +2ab(XY - XE(Y) - YE(X) + E(X)E(Y)] \\
&= E[a^2(X-E(X))^2 + b^2(Y-E(Y))^2 + 2ab(X-E(X))(Y-E(Y))] \\
&= a^2E[(X-E(X))^2] + b^2E[(Y-E(Y))^2] + 2abE[(X-E(X))(Y-E(Y))] \\
&= a^2\mathrm{Var}(X) + v^2\mathrm{Var}(Y) + 2abE[(X-E(X))(Y-E(Y))].
\end{aligned}$$

We isolate the right-most term in the last line and give it a name.

Definition. Let X and Y be random variables. The *covariance*, $\mathrm{Cov}(X,Y)$, is

$$\mathrm{Cov}(X,Y) = E[(X - E(X))(Y - E(Y)] = E[XY] - E(X)E(Y).$$

The *correlation coefficient*, $\rho_{X,Y}$, is

$$\rho_{X,Y} = \frac{\mathrm{Var}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

If $\rho_{X,Y} = 0$ then X and Y are *uncorrelated*.

Note. You are likely familiar with the graphical interpretation of the correlation coefficient, as considered in Foundations of Probability and Statistics-Calculus Based (MATH 2050) in Section 2.6. Jointly Distributed Random Variables, and even in Probability and Statistics-Noncalculus (MATH 1530) in Chapter 4. Scatterplots and Correlation. In Mathematical Statistics 1 (STAT 4047/5047) in Section 2.5. The Correlation Coefficient, it is proved that $-1 \leq \rho_{X,Y} \leq 1$ (see Theorem 2.5.1) and that for independent X and Y we have $\rho_{X,Y} = 0$ (see Theorem 2.5.2). Gut comments (see page 10): "There is a famous result to the effect that two independent random variables are uncorrelated but that the converse does not necessarily hold." In the same Mathematical Statistics 1 notes, Example 2.5.3 addresses the converse.

Revised: 10/9/2022